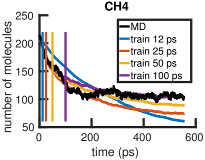
Statistical Learning of Reduced Kinetic Monte Carlo Models of Complex Chemistry from Molecular Dynamics

Qian Yang, Carlos A. Sing-Long, Enze Chen, Muralikrishna Raju, Evan J. Reed
Stanford University

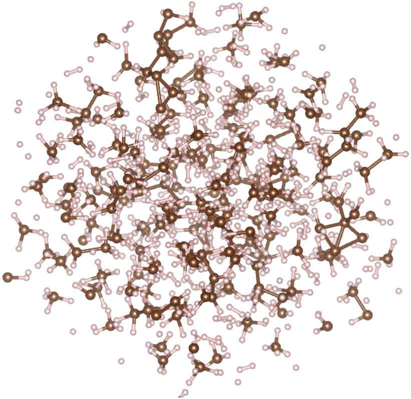


February 26, 2018

AI at SLAC Seminar

Molecular dynamics simulations (existing or new) contain a wealth of data.

- 1000s of atoms
- 1000s of reactions
- 100s of molecules
- 1000s of timesteps per picosecond simulation



Can we do more with this complex, expensive DATA?

We propose using statistical learning to learn a chemical reaction network from atomistic simulation data.

- We discover that it's possible to build kinetic Monte Carlo (KMC) models from a single or few MD simulations that can extrapolate the dynamics of the chemical system **more than 10x in time** and in **chemical space**.

- We develop a new data-driven method that **reduces thousands of reactions** to fewer than a hundred in a matter of minutes.

Outline

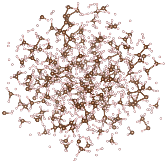
Building KMC models from MD simulations

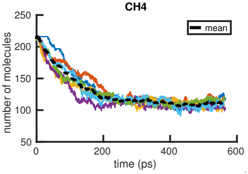
Extrapolation in Time and Chemical Space

Reduction from thousands to dozens of reactions

Summary

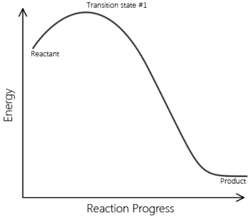
We project molecular dynamics data onto a model consisting of elementary reactions and reaction rates.



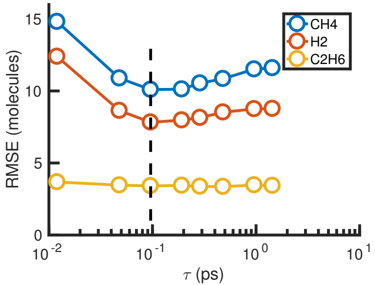


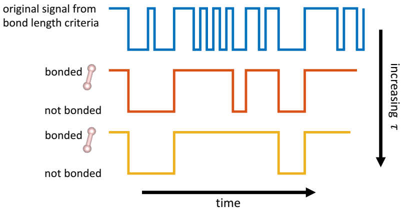
⇓

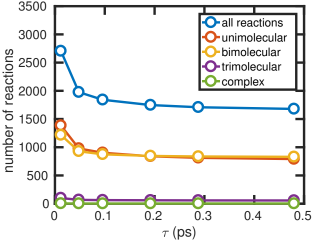
Elementary Reactions and Reaction Rates



Bond duration is used to control bias-variance tradeoff in our model







We model chemical reaction networks as a set of elementary reactions and reaction rates governed by the chemical master equation.

Reactions are random events. We associate with every reaction a **propensity function**, $a_j(x)$, such that $a_j(x)dt$ gives the probability of that reaction occurring in the time interval $[t, t+dt)$.

- Unimolecular reactions $X_1 \to products$: $a_j(\mathbf{X}) = k_j X_1$
- Bimolecular reactions $2X_1 \to products$: $a_j(\mathbf{X}) = k_j X_1(X_1 - 1)$
- Bimolecular reactions $X_1 + X_2 \to products$: $a_j(\mathbf{X}) = k_j X_1 X_2$

The **chemical master equation** is a system of ODEs that gives the probability $\mathbb{P}(x, t)$ of being in a particular state $\mathbf{X}(t) = x$ in molecular concentration space at time t.

We use Maximum Likelihood Estimation to estimate reaction rates

In a given MD simulation with M unique species and R unique reactions, we observe at every timestep t:

1. vector of molecule counts $\mathbf{X}(t) = (x_1(t), ..., x_M(t))$
2. vector of reaction counts $\mathbf{R}(t) = (r_1(t), ..., r_R(t))$

For $t \geq 1$, we can consider $\mathbf{X}(t)$ as a function of $\mathbf{X}(0)$ and all of the previous $\mathbf{R}(t')$ for $t' < t$, so that our set of observations for timesteps 1 to T are

$$(\mathbf{X}(0), \mathbf{R}(0), ..., \mathbf{R}(T-1))$$

For N independent MD simulations, we estimate the vector of reaction rate constants $\mathbf{k} = (k_1, ..., k_R)$ by maximizing the likelihood of the observations:

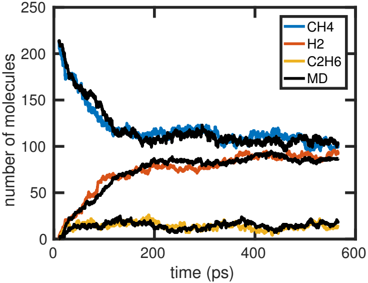
$$\mathbf{k}^* = \max_{\mathbf{k}} \ \mathcal{L}(\mathbf{k}; (\mathbf{X}(0), \mathbf{R}(0), ..., \mathbf{R}(T-1))_1, ..., (\mathbf{X}(0), \mathbf{R}(0), ..., \mathbf{R}(T-1))_N)$$

We simulate the chemical master equation using kinetic Monte Carlo

Given a reaction network with thousands of reactions and reaction rates, Gillespie Stochastic Simulation is equivalent to kinetic Monte Carlo and can be used to simulate the chemical master equation exactly.

At each step:

- Choose the **next reaction** based on the propensity functions for each reaction.
- Choose the **time** until the next reaction.



Our KMC simulations take minutes to model the MD data that took weeks to generate.

Outline

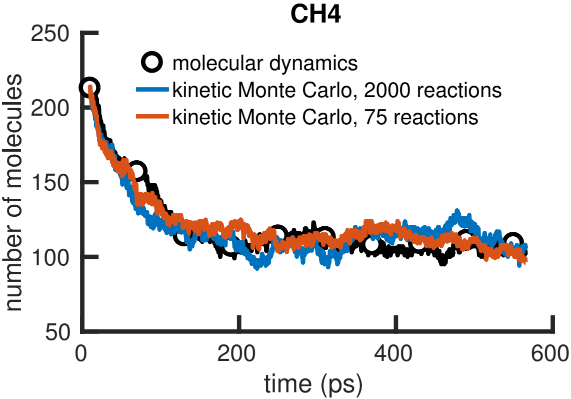
Building KMC models from MD simulations

Extrapolation in Time and Chemical Space

Reduction from thousands to dozens of reactions

Summary

We show that the chemical reaction network can be significantly reduced when describing target molecules.



Model reduction of complex systems is a formidable technical challenge:

- Sensitivity Analysis
- Mixed Integer Programming

We represent chemical reaction networks as a linear system in the set of reactions

Consider the toy reaction network:

$$A + B \to C \tag{1}$$
$$A \to 2C \tag{2}$$
$$2C \to A \tag{3}$$

Then according to the reaction rate equations, we have

$$\begin{aligned}
\frac{d[A]}{dt} &= -k_1[A][B] - k_2[A] + k_3[C]\left([C] - 1\right) \\
\frac{d[B]}{dt} &= -k_1[A][B] \\
\frac{d[C]}{dt} &= k_1[A][B] + 2k_2[A] - 2k_3[C]\left([C] - 1\right)
\end{aligned}$$

We would like to choose reaction rate coefficients to match the first and second moments of the stochastic model at the sampled $\mathbf{X}(t)$. The moments are **nonlinear** in molecular concentrations, but **linear** in reaction rate coefficients.

We represent chemical reaction networks as a linear system in the set of reactions

$$\frac{1}{\tau}\mathbb{E}\underbrace{\begin{bmatrix} A_{t+\tau}-A_t \\ B_{t+\tau}-B_t \\ C_{t+\tau}-C_t \end{bmatrix}}_{y_t} = \mu_{t+\tau}$$

$$= k_1 \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix} A_t B_t + k_2 \begin{bmatrix} -1 \\ 0 \\ 2 \end{bmatrix} A_t + k_3 \begin{bmatrix} 1 \\ 0 \\ -2 \end{bmatrix} C_t\left(C_t - 1\right)$$

$$= \underbrace{\begin{bmatrix} -1 & -1 & 1 \\ -1 & 0 & 0 \\ 1 & 2 & -2 \end{bmatrix}}_{\mathbf{R}} \underbrace{\begin{bmatrix} A_t B_t & 0 & 0 \\ 0 & A_t & 0 \\ 0 & 0 & C_t\left(C_t-1\right) \end{bmatrix}}_{\mathbf{D}_t} \underbrace{\begin{bmatrix} k_1 \\ k_2 \\ k_3 \end{bmatrix}}_{\mathbf{k}}$$

A similar expression linear in the vector of k_i's can be obtained for

$$\frac{1}{\tau}\mathsf{cov}\left(y_t\right) = \mathbf{H}\mathbf{D}_t\mathbf{k}$$

We represent model reduction of chemical reaction networks as a linear system subject to L1 regularization

$$\frac{1}{\tau}\begin{bmatrix} \mu_2|_{X_1} \\ \Sigma_2|_{X_1} \\ \dots \\ \mu_{n+1}|_{X_n} \\ \Sigma_{n+1}|X_n \end{bmatrix} = \begin{bmatrix} RD_1 \\ HD_t \\ \dots \\ RD_n \\ HD_n \end{bmatrix}\mathbf{k} \qquad \leftarrow \textbf{sample nonlinear data}$$

Then after scaling the system by **k** from the MLE, we have the nonnegative LASSO (least absolute selection and shrinkage operator):

$$\min_{\mathbf{c}} \|A\mathbf{c} - b\|_2$$

$$\text{s.t.} \|\mathbf{c}\|_1 \leq \lambda, \quad \mathbf{c} \geq 0$$